or, more conveniently for digital computation,

$$p^+_{i+1,j}=R_{i+1}(p^+_{i,j-1}-p^-_{i+1,j-1})+p^+_{i,j-1}$$
$$p^-_{i,j}=R_{i+1}(p^+_{i,j-1}-p^-_{i+1,j-1})+p^-_{i+1,j-1}. \quad (6.89)$$

The last pipe element of the line terminates in a load that is the radiation impedance of the mouth. Let A_L be the area of the last pipe element and Z_L the terminating radiation load. At the load terminals (the end of pipe A_L), the right-going and left-going pressure waves satisfy

$$\frac{p_L}{U_L}=Z_L=\frac{A_L(p^+_L+p^-_L)}{\rho c(p^+_L-p^-_L)}.$$

If Z_L is written in terms of a z-transform, the reflected wave p^-_L can be obtained in terms of weighted and delayed values of p^+_L; that is, a reflection coefficient relation can be set down in which $p^-_L=p^+_L f(z^{-1})$. The load pressure $(p^+_L+p^-_L)$ produces a mouth volume velocity U_L through Z_L, which, when differentiated, represents the radiated pressure. Formulations such as these have been used in a phoneme-driven vocal-tract synthesizer (KELLY and LOCHBAUM) and in a simulation of articulatory activity (MERMELSTEIN).

A further useful approach for digital representation of the distributed vocal tract follows the bilateral transmission line developed in Chapter III. Once the line composed of elemental T or π sections is set down, transient solutions of pressure and volume velcoity along the line may be obtained from difference equation approximations to the differential equations for the network. Area variations are reflected in network element changes. This approach also permits duplication of absolute acoustic impedances. For this reason it has been used in a vocal-tract synthesizer to study the acoustics of vocal-cord vibration and turbulent sound generation (FLANAGAN and CHERRY; see Section 6.243).

VII. Perception of Speech and Speech-Like Sounds

As a general topic, auditory perception can be divided a number of ways. From the standpoint of communication, one separation might be between classical auditory psychophysics, on the one hand, and the recognition of acoustic signals presented within a linguistic framework, on the other. The former relates principally to the abilities and limitations of the hearing organ as a mechano-neural transducer of all acoustic signals. The latter bears mainly upon the identification and classification of auditory patterns which are significant within the communicative experience of the listener.

Classical auditory psychophysics strives to discover the "resolving power" of the hearing mechanism. Discrimination is usually examined along fundamental dimensions of the stimulus—usually along only one dimension at a time. The measurements are generally conducted under conditions which are most favorable for making the relevant discriminations, that is, differential discriminations or close comparisons. Differential thresholds for dimensions such as intensity and frequency fall into this classification. Intuitively one feels that large neural storage and complex central processing probably are not brought into play in such detections. The measures more likely reflect the capacity of the transducer and the peripheral neural net to preserve details about a given stimulus dimension. The discussion in Chapter IV, for example, touched upon these properties of the peripheral system. The apparent relations between physiological and psychoacoustic response were analyzed for several stimulus cases. The acoustic signals were of the "classical" type in that they were either temporally punctate or spectrally simple, or both.

Speech, on the other hand, is a multidimensional signal that elicits a linguistic association. For it to be an effective communication code, some sort of absolute perceptual categorization must be made of its content. That is, the signal must be broken down into a finite number of discrete message elements. The "size" of these perceptual elements, and the manner in which they are processed to yield the percept, are questions of considerable debate and not little speculation. Our present knowledge brings us nowhere near a good understanding of the process. Theorizing about speech perception—cloaked in all of its linguistic and over-learned functions—abounds with pitfalls. An even larger problem, perhaps, is reconciling physiological, psychophysical and linguistic factors. As in other difficult situations, it is tempting to push back to some still-higher center the final decision-making process that is the real seat of perception.

Although a complete theory of speech perception remains in the future, a good deal can be said about auditory discrimination. Some of the "classical" measurements relate strongly to signal dimensions important to speech—even though the measurements are made outside of linguistic or contextual frames. In addition, a respectable amount of information has been accumulated on the acoustic cues associated with synthetic approximants to simple speech elements—for example, syllables and phonemes.

From the practical point of view, articulation tests and intelligibility measures based upon absolute recognition of sentences, words, syllables, and isolated phonemes can be used to good effect in evaluating transmission facilities. For a given processing of the voice signal, these tests often help to identify factors upon which perception depends (although

they serve poorly, if at all, in supplying a description of the perception process itself). Under certain conditions, the so-called articulation index can be used to compute intelligibility scores from physical measurements on the transmission system. Still ancillary to intelligibility testing, some data are available on the influences of linguistic, contextual and grammatical constraints. Contrariwise, measures of the prosodic and quality features of speech are not well established.

The present chapter proposes to circumscribe some of these problems. In particular, the aim is to indicate current level of understanding in the perception of speech and speech-related sounds.

7.1. Differential vs. Absolute Discrimination

Classical psychophysics generally deals with discriminations made in close comparison. Speech perception, on the other hand, seems more likely an absolute classification of an acoustic signal. Can the former provide useful information about the latter, or vice versa?

Man is highly sensitive to differences in the frequency or intensity of sounds presented for comparison. Under certain conditions the threshold for detecting a difference in the frequencies of two successively presented pure tones may be as small as one part in 1000 (ROSENBLITH and STEVENS). The threshold for detecting a difference in intensity may be less than one db (RIESZ). On the basis of comparative judgments, it has been estimated that the normal listener can distinguish about 350000 different tones (STEVENS and DAVIS).

Contrasting with this acute differential sensitivity, a listener is relatively inept at identifying and labelling sounds presented in isolation. When equally-loud pure tones are presented individually for absolute judgment of frequency, listeners are able to accomplish perfect identification among only five different tones (POLLACK). This identification corresponds to an information transfer of about 2.3 bits per stimulus presentation. If, however, the sound stimulus is made multidimensional, for example by quantizing it in frequency, loudness, duration, etc., the correct identifications increase, and the information transferred may be as high as five to seven bits per stimulus presentation (POLLACK and FICKS). This rate is equivalent to correct identification from an ensemble of from 32 to 128 stimuli.

It is clear that absolute and differential discriminations yield substantially different estimates of man's informational capacity. The former suggest that few subdivisions along a given dimension can be identified, whereas the latter indicate that a much larger number may be discriminated. The differential measure, however, reflects the ability to discriminate under the most favorable circumstances for detecting a difference (namely, a close comparison usually along a single dimension). In a sense, it represents an upper bound on the resolving ability of the perceptual mechanism.

So far as absolute judgments are concerned, the differential estimate of discrimination is an optimistic one. The probability is extremely small that stimulus quantizations as small as a difference limen[1] could ever be detected absolutely. Even so, differential measures quantify perception in a "clinical" way, and they place a rough ceiling (albeit over-optimistic) on the ability to detect changes in a signal. In any speech processing system, fidelity criteria based upon differential discriminations would be expected to be very conservative. Lacking more directly applicable measures, however, they can often be useful in estimating the performance and channel capacity requirements of a transmission system (FLANAGAN, 1956b).

7.2. Differential Discriminations Along Signal Dimensions Related to Speech

The results of Chapter III and IV suggest that significant dimensions for the speech signal might be defined in either the acoustic or articulatory domains. Both domains have perceptual correlates. The analyses in Chapter III, for example, attempted to separate the properties of vocal transmission and excitation. Perceptually-important acoustic dimensions of the system function are those of the mode pattern—that is, the complex frequencies of the poles and zeros of the transmission. Alternatively, the same information is specified by the bandwidths and frequencies of the maxima and minima of the amplitude spectrum and by the values of the phase spectrum. In a similar manner, relevant dimensions of the excitation source for voiced sounds are intensity, fundamental frequency and perhaps spectral zero pattern (or equivalently, glottal wave asymmetry and duty factor). For the unvoiced source, intensity and duration are significant dimensions.

Auditory sensitivity to some of these factors—divorced of any linguistic or contextual frame—has been measured in psychoacoustic experiments. For example, data are available on just-discriminable changes in formant frequency, fundamental frequency, over-all intensity and formant bandwidth. Without going into the details of any specific experiment, the nature of the results can be summarized.

7.21. Limens for Vowel Formant Frequencies

Using synthetic vowel sounds generated by a terminal-analog synthesizer (see Section 6.22, Chapter VI), just-discriminaable changes in

[1] The terms difference limen (DL) and just-noticeable difference (JND) are synonomous with differential threshold or just-discriminable change.

the frequencies of the first and second formants have been measured (FLANAGAN, 1955b). The synthesizer was operated as it would be used in a formant-vocoder system. Although the difference limens (DL's) depend to an important extent upon the proximity of the formants, they are found to be on the order of three to five percent of the formant frequency[1].

7.22. Limens for Formant Amplitude

The results of Chapter III and VI show that the relative amplitude of a given formant in the speech signal is a function of several factors, among them formant frequency, vocal damping, transmission zeros and excitation characteristics. One measure of the differential discriminability of formant amplitude has been made with a parallel-connected, terminal-analog synthesizer (FLANAGAN, 1957a). The intensity limen for the second formant of a near-neutral vowel (/æ/) is found to be about 3 db.

A related measurement of the limen for over-all intensity of a synthetic vowel gives a value of about 1.5 db (FLANAGAN, 1955a). Because the first formant is usually the most intense formant in vowel sounds, the over-all figure might be taken as a rough estimate of the first-formant intensity limen.

Another experiment determined the intensity limens for single harmonic components of synthetic vowels (FLANAGAN, 1965). Values found for intensity changes at the first and second formant frequencies support well the values just mentioned. Intensity limens for harmonic components located in spectral "valleys" can be quite large, as much as +13 db to $-\infty$ db, i.e., complete absence.

7.23. Limens for Formant Bandwidth

Apparently, no direct measures of the discriminability of changes in formant bandwidth, or damping, have been made on synthetic vowels. However, some related measurements, and their extrapolations, suggest what might be expected.

STEVENS (1952) measured the descriminability of changes in the tuning and damping of a single electrical resonator. The resonator was excited by periodic pulses at a fundamental frequency of 125 cps. The output signal was therefore representative of a one-formant vowel. In

[1] This experiment considered changes in the frequency of only one formant at a time. In real speech—and in formant-coding of speech—the formants usually move simultaneously. A relevant and practically-useful extention of the experiment might be the determination of "DL solids" in $F1-F2-F3$ space. Proximity effects of the formants should, in general, give these "solids" ellipsoidal shapes. Similar comments about discrimination of simultaneous changes in signal dimensions apply in several of the following experiments.

general, changes on the order of 20 to 40% in formant bandwidth were just-discriminable.

Also, the results of Chapter III show that the amplitude of a formant peak is inversely related to formant damping. The 1.5 db figure found for the amplitude limen of the first formant corresponds to a bandwidth change of about 20%. Similarly, the 3 db figure for the second formant corresponds to a bandwidth change of about 40%[1].

7.24. Limens for Fundamental Frequency

Following an experimental procedure similar to that used with the formant measurements, a difference limen has been measured for the fundamental excitation frequency of synthetic vowel sounds (FLANAGAN and SASLOW). For synthetic vowels appropriate to a man, the fundamental-frequency limen is found to be about 0.3 to 0.5 per cent of the fundamental-frequency. From this and the previously-mentioned measurements, a hierarchy in frequency acuity emerges. The formant-frequency limen is an order of magnitude more acute than the formant-bandwidth limen, and the fundamental-frequency limen is an order of magnitude more acute than the formant-frequency limen.

7.25. Limens for Excitation Intensity

For a given glottal wave shape and vocal transmission, the over-all intensity of a voiced sound is directly proportional to the amplitude of the glottal pulse. As mentioned previously, a measure of the limen for over-all vowel intensity gives a value of about 1.5 db.

Similarly, the over-all intensity of an unvoiced sound is directly related to the amplitude of the unvoiced source. Fricative consonants are relatively broadband, noise-excited, continuant sounds. The discriminability of changes in their over-all amplitude might be expected to be somewhat similar to that of wide-band noise. Intensity limens have been measured for the latter (G. A. MILLER, 1947). They are found to be of the order of 0.4 db for sensation levels above 30 db. The minimum perceptible intensity change is therefore about 5%. Only a few fricative consonants have relatively flat spectra, but the figure might be used as an order-of-magnitude estimate. Experience with speech synthesis confirms that it is a conservative figure.

7.26. Limens for Glottal Zeros

The differential discriminability of changes in the spectral zero pattern of the vocal cord source (see Section 6.24, Chapter VI), or in the

[1] Another multidimensional DL of interest might be that for simultaneous changes in formant bandwidth and frequency. In other words, one might determine DL "areas" in the complex-frequency plane for vocal tract poles.

detailed spectrum of the glottal wave, have, to the author's knowledge, been observed only informally (FLANAGAN, 1961 b). The glottal source may contribute significant factors to speech quality and to speaker recognition. Therefore, liminal measures of parameters such as the duty factor and asymmetry of the glottal wave could be valuable in establishing bounds on their importance to speech naturalness.

It is clear that if complex source zeros lie far enough away from the $j\omega$-axis of the frequency plane they have negligible effect on signal quality. One experiment in which only gross features of the source spectrum and waveform were preserved suggests that many temporal and spectral details are unimportant to quality (ROSENBERG) (see Section 6.241 h).

7.27. Discriminability of Maxima and Minima in a Noise Spectrum

The vocal tract transmission for fricative consonants, like other sounds, is characterized by certain poles and zeros. Broadband noise excitation is filtered by this transmission. Some of the poles and zeros (and their related spectral maxima and minima) are perceptually significant. Others are not. One measurement considered the differential discriminability of a single peak or valley in an otherwise flat noise spectrum (MALME). A single pole and zero filtering of a broadband noise was used to produce the spectral variations shown in the insert of Fig. 7.1. The equivalent complex frequencies (half-power bandwidths vs. center frequencies) of the irregularities which were just-detectable from the flat spectrum are also plotted in Fig. 7.1. The db numbers next to the points are the just-perceptible peak heights and notch depths, respectively. These data indicate that, at least in a flat-noise surround, spectral peaks with Q's (i.e., ratios of center frequency to bandwidth) less than about 5, and spectral notches with Q's less than about 8 are not differentially perceptible.

The results suggest, therefore, that many of the small spectral irregularities seen in a fricative consonant such as /f/ are not perceptually significant. In the same vein, certain spectral peaks such as in /s/ or /ʃ/ are of course significantly different from a flat spectrum. Synthesis of fricative consonants has been demonstrated by representing the spectrum in terms of two poles and one zero (HEINZ and STEVENS). Appropriate Q's for the poles are found in the range of about 5 to 13. For the zero, Q's of the order of 2 to 4 appear appropriate. The suggestion is, therefore, that to the extent the results in Fig. 7.1 can be applied, the poles are more significant perceptually than the zero, the latter apparently having importance only in contributing to the gross spectral shape.

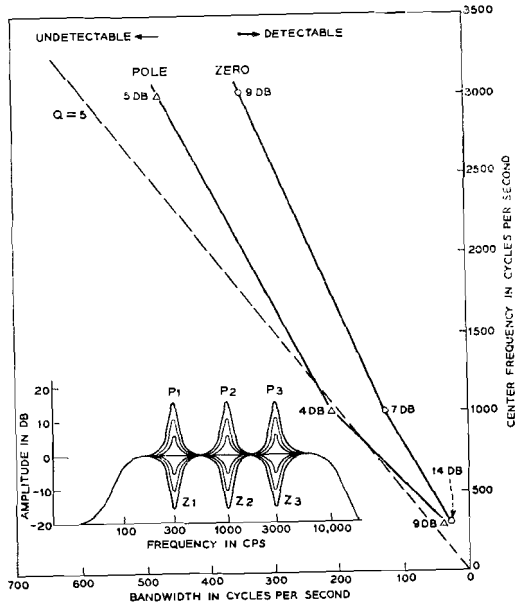


Fig. 7.1. Detectability of irregularities in a broadband noise spectrum. (After MALME)

This appears to be the case, for the zero has been found to be relatively noncritical of frequency position and often can be placed automatically about an octave below the first pole (HEINZ and STEVENS).

A similar discrimination measurement has been made for a noise spectrum with exactly periodic maxima—that is, for a comb filtering of the noise (ATAL and SCHROEDER). The objective was to investigate the preceptual effects of irregularities in the frequency response of rooms. The comb-filtered noise was differentially compared with white noise of equal power, and a limen was obtained for the minimum detectable periodic irregularity. The minimum detectable ratio of maximum-to-minimum spectral amplitude was found to be about one db. This figure is in close agreement with the intensity limen measured for white noise (see Section 7.25).

The results of this same experiment provide information on the weighting function used by the ear in performing its short-time spectral analysis. The weighting function deduced from the measurements is approximately exponential in form, with an initial slope corresponding to a time constant of 9 msec. This latter figure compares favorably with the time constant deduced for loudness measurements on periodic clicks (see Section 4.33, Chapter IV).

7.28. Other Close-Comparison Measures Related to Speech

A number of other psychophysical measurements relate, more or less strongly, to differential perception along speech dimensions. Several of these can be mentioned to illustrate the diverse nature of the data.

One experiment measured the perception of a single, time-varying formant (BRADY, HOUSE, and STEVENS). A continuously-tunable resonant circuit was excited by five equi-spaced pitch pulses. The pulses were produced at a rate of 100 sec^{-1}. During the excitation, the tuning of the resonator was moved between 1000 and 1500 cps, according to the rising and falling trajectories shown in Fig. 7.2. The formant transitions were accomplished in 20 msec. To examine how the varying formant frequency is perceived, listeners were asked to adjust the frequency of a nontime-varying resonance until it sounded as much like the varying one as possible. Typical results of the matches are shown in Fig. 7.3. The test stimuli labelled *a*, *b*, *c*, *d*, *e*, and *f* correspond to those diagramed in Fig. 7.2. The data show a strong tendency to set the steady resonance to a frequency corresponding to the final value of the varying formant, particularly when the formant change occurs near the beginning of the sound. The tendency to match the final frequency appears somewhat stronger for stimuli in which the resonant frequency ascends.

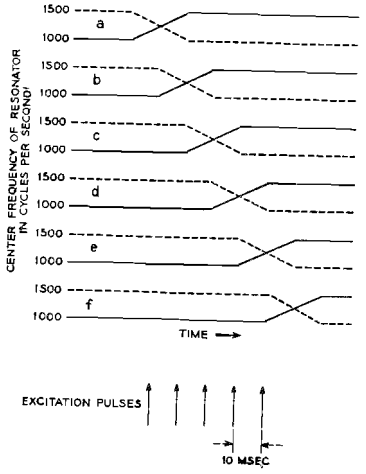


Fig. 7.2. Frequency paths and excitation pattern for a simulated time-varying formant. Rising and falling resonances are used. The epochs of the five excitation pulses are shown. (After BRADY, HOUSE and STEVENS)

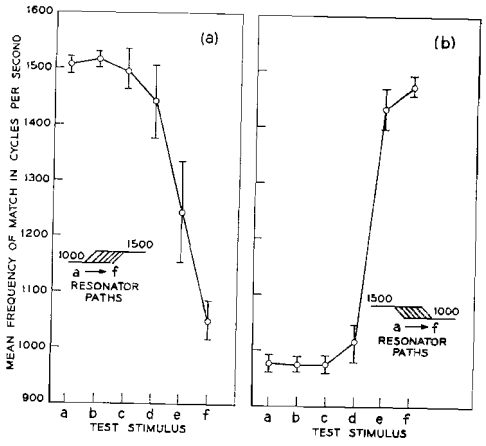


Fig. 7.3 a and b. Results of matching a nontime-varying resonance to the time-varying resonances shown in Fig. 7.2. Mean values are plotted. The vertical lines indicate the standard deviations of the matches. (After BRADY, HOUSE and STEVENS)

In a different vein, a temporal fine structure is known to exist in the glottal source. The shape and periodicity of the glottal pulse is subject to various perturbations which may be significant to speech quality. In a condition known as diplophonia, for example, alternate glottal pulses may be of different size (S. SMITH). Similarly, successive glottal periods may vary in duration. To quantify this effect, one study analyzed the durations of 7000 pitch periods of real speech (LIEBERMAN). In successive samples of three periods each, the variation in period was greater than ± 0.1 msec in 86% of the cases. In 20% of the cases the duration difference between periods was greater than 0.6 msec, and in 15% it was greater than 1.0 msec. In 38% of the cases the periods were alternately long and short. Adjacent periods were not correlated, but alternate periods were highly correlated.

As one step toward trying to understand the possible perceptual correlates of these factors, a preliminary investigation has examined the effects upon perceived pitch of systematic differences in amplitude and timing of an otherwise periodic pulse train (FLANAGAN, GUTTMAN, and WATSON; GUTTMAN and FLANAGAN, 1962). Among the conditions considered were the pulse wave forms shown in the left-hand column of Fig. 7.4. Starting with exactly periodic trains (of period $T/2$), alternate

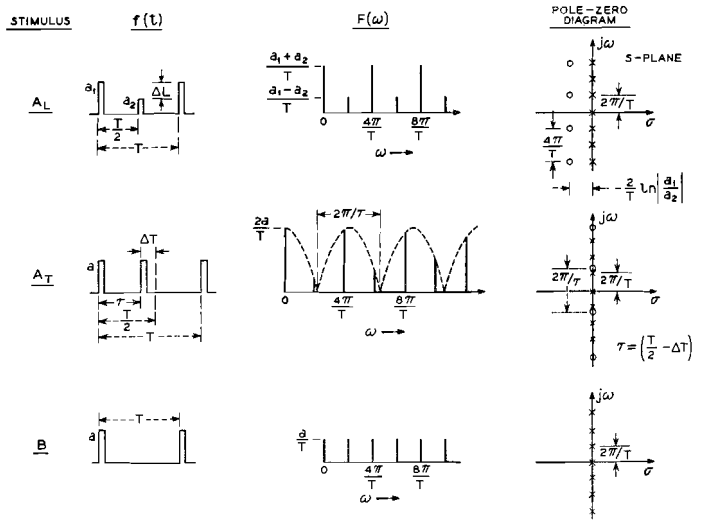


Fig. 7.4. Periodic pulse stimuli for assessing the influence of amplitude and time perturbations upon perceived pitch. The left column shows the time waveforms of the experimental trains; amplitude variation (A_L), time variation (A_T), and the standard matching train (B). The second column shows the corresponding amplitude spectra, and the third column shows the complex-frequency diagram. (After FLANAGAN, GUTTMAN and WATSON; GUTTMAN and FLANAGAN, 1962)

pulses in the train were changed incrementally either in amplitude level (Stimulus A_L) or in time of occurrence (Stimulus A_T). The effect upon pitch was assessed by having listeners adjust the frequency of a uniform, periodic train (Stimulus B) until its pitch matched that of the perturbed train. As either the amplitude difference (ΔL) or the time difference (ΔT) increases, a point is soon reached where the pitch drops by an octave.

The second column of Fig. 7.4 shows the frequency spectra of the amplitude-varied stimulus (A_L), the time-varied stimulus (A_T), and the standard matching stimulus (B). The third column of the figure shows the corresponding pole-zero diagrams for the three periodic trains. Notice that for the A_L signal the relative amplitudes of adjacent spectral lines are dependent only upon the pulse amplitudes a_1 and a_2. For A_T, on the other hand, the spectral amplitudes are conditioned by the fundamental period T and by the cycloidal envelope which, in turn, is determined by the interval τ.

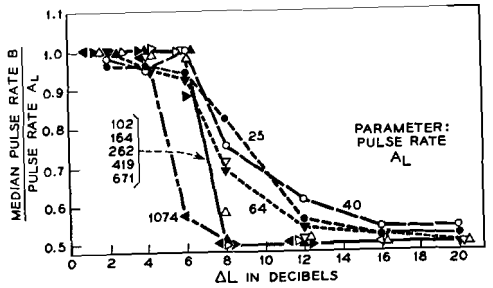
Median matches made by a number of listeners for the ΔL and ΔT conditions are shown in Figs. 7.5a and 7.5b, respectively. In both plots the parameter is the pulse rate of the A stimulus (i.e., twice its fundamental frequency). The results in Fig. 7.5a for ΔL show that, over the frequency range appropriate to the human voice fundamental, an amplitude difference ΔL of about 6 to 8 db, or greater, will produce an octave reduction in the perceived pitch. In the same fashion, the ΔT data in Fig. 7.5b show that in the range of the voice fundamental (i.e., about 100 pps and above) a time shift $2\Delta T/T$ on the order of 0.1 or more, will produce an octave reduction in pitch.

7.29. Differential Discriminations in the Articulatory Domain

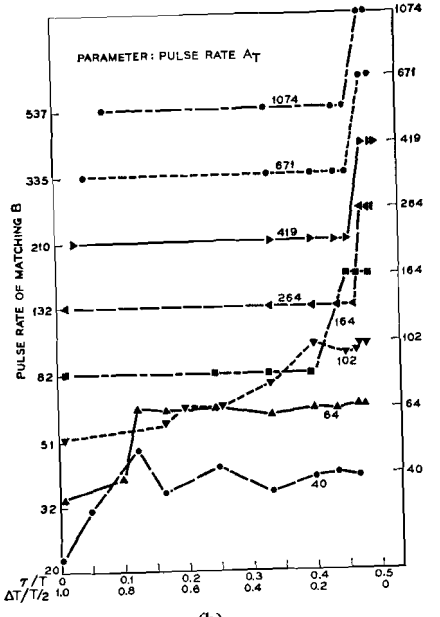
The acoustic dimensions considered for the speech and speech-like signals in the preceding discussion have counterparts in the articulatory domain. However, the acoustic and articulatory relations do not generally exist in one-to-one correspondence. For example, a change in a constriction size, or in its location, alters not only one formant frequency, but in general all of them (see Fig. 3.39, Chapter III). It is therefore difficult to interpret, say, limens for formant frequency and amplitude in terms of just-discriminable articulatory changes. One can, nevertheless, make some simple observations about the links between the domains.

The just-discriminable changes in formant frequency were found to be about three to five per cent. For a straight pipe the formants are approximately

$$F_n=\frac{(2n-1)c}{4l}, \quad n=1,2,\ldots.$$



(a)



(b)

Fig. 7.5a and b. Results of matching the pitch of a uniform pulse train (B) to that of: (a) a periodic train (A_L) whose alternate pulses differ in amplitude by ΔL; and (b) a periodic train (A_T) whose alternate pulses are shifted in time by ΔT. In both cases the parameter is the pulse rate of the A stimulus. (After FLANAGAN, GUTTMAN and WATSON; GUTTMAN and FLANAGAN, 1962)

The sensitivity of the mode frequencies to length changes is

$$\partial F_n/\partial l=-\frac{(2n-1)c}{4l^2}, \quad \text{or} \quad \frac{F_n}{\Delta F_n}=-\frac{l}{\Delta l},$$

so that a given percentage change in the tract length l produces the same percentage change in the formant frequencies. The DL for tract length

might therefore be expected to be roughly comparable, percentage-wise, to the formant frequency *DL*. By referring to Fig. 3.39, Chapter III, one can see other, more complex correspondences between formant changes and articulatory changes.

Another simple example is the sensitivity of the mode damping for a straight pipe to changes in the mean glottal area (see Eq. (3.74)]. Assume for simplicity that the equivalent glottal impedance is purely resistive and is produced only by kinetic factors, that is,

$$R_g' = \frac{(2\rho P_{s_0})^{\frac{1}{2}}}{A_0}$$

[using the notation of see Eq. (3.51)]. The pole dampings (i.e., real parts) are given by

$$\sigma_n \cong -\left(\alpha c + \frac{Z_0 c}{l R_g}\right),$$

or

$$\sigma_n \cong -\left[\alpha c + \frac{c Z_0 A_0}{l(2\rho P_{s_0})^{\frac{1}{2}}}\right]$$

[see Eq. (3.74)]. The sensitivity of the damping with respect to mean glottal area is then

$$\frac{\partial \sigma_n}{\partial A_0} \cong -\frac{c Z_0}{l(2\rho P_{s_0})^{\frac{1}{2}}},$$

or the change in mode damping is approximately proportional to the change in mean glottal area.

7.3. Absolute Discrimination of Speech and Speech-Like Sounds

Most efforts to establish the acoustic cues for speech-sound recognition have been absolute identification experiments. The test stimuli have generally been synthetic versions of phoneme-length and syllable-length utterances. This approach presumably keeps the stimuli simplified to noncontextual situations where only the physical properties of the specific signal influence the percept. At the same time it may permit association of a linguistic structure, and the perceptual responses are usually interpreted within this frame of reference.

7.31. Absolute Identification of Phonemes

A relatively small number of experiments has dealt solely with isolated phonemes. One study—using a transmission-line vocal tract analog—investigated articulatory configurations appropriate to vowels. It

tested a simple three-number articulatory description of vowel production (Stevens and House, 1955; House and Stevens, 1956). The three-number scheme for describing vowel articulation is illustrated for two configurations in Fig. 7.6. The three parameters used to describe the vocal shape are the radius of the maximum constriction, r_0; the distance from the glottis to the constriction, x_0; and the ratio of mouth area to lip rounding, A/l. The radius of the dashed portion of the tract is described by the function

$$r(x) = [0.025(1.2 - r_0)(x - x_0)^2 + r_0],$$

where the lengths are in centimeters.

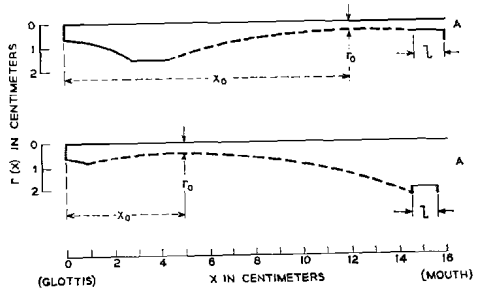


Fig. 7.6. Three-parameter description of vowel articulation. r_0 is the radius of the maximum constriction; x_0 is the distance from the glottis to the maximum constriction; and A/l is the ratio of mouth area to lip rounding. (After Stevens and House, 1955)

An electrical transmission line simulated the configurations and synthesized the sounds. Isolated vowels, 500 msec in duration, were judged absolutely by listeners and placed into nine English vowel categories. Pitch was monotonically inflected from 120 to 140 cps. The listener responses in terms of articulatory parameters are illustrated for one value of constriction in Fig. 7.7. The two response contours indicate agreement among 50% and 75% of the responses, respectively. The Peterson and Barney data for natural vowels uttered by men (see Fig. 5.10, Chapter V), when transformed into the same articulatory coordinates, are given in Fig. 7.8. The two plots show that, except for small differences, the three number description does surprisingly well in providing a unique specification of the vowels.

A somewhat similar experiment on synthesis and perception has been carried out for Japanese vowels (Nakata and Suzuki). In this experiment, however, the sounds were produced by a terminal-analog synthesizer, and the idea was to find the synthetic formant patterns appropriate to the vowels.

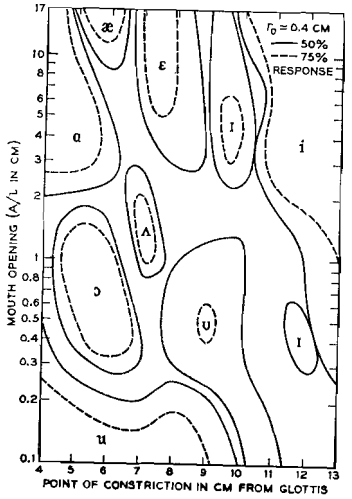


Fig. 7.7. Listener responses to isolated synthetic vowels described by the 3-parameter technique. One value of constriction is shown. Two levels of response corresponding to 50 and 75 % agreement among subjects are plotted. (After HOUSE and STEVENS, 1955)

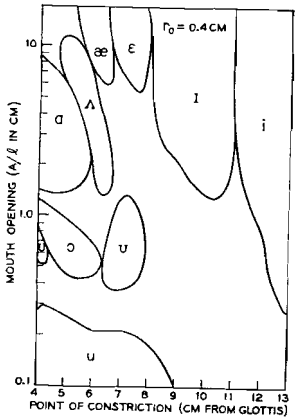


Fig. 7.8. Formant frequency data of PETERSON and BARNEY for 33 men transformed into the 3-parameter description of vowel articulation. (After HOUSE and STEVENS, 1955)

The same transmission-line analog – but with attached nasal tract – has been used to study the perception of nasal consonants (HOUSE). Isolated, 500 msec representations of nasal consonants were synthesized and presented to listeners for absolute judgment. The permissible response categories were the three nasal consonants /m, n, and η/. The articulatory description used for synthesis was similar to that described in the preceding discussion on vowels, but with the additional specification of the velar coupling. Typical confusion matrices of responses (to articulatory configurations which were determined by pre-tests to be representative nasal consonant stimuli) are shown in Table 7.1 a.

Table 7.1. *Listener responses to synthetic and natural nasal consonants*

a) Synthetic				b) Natural			
Stimulus	Response %			Stimulus	Response %		
	m	n	ŋ		m	n	ŋ
m	81	11	8	m	96	4	0
n	33	61	6	n	42	56	2
ŋ	20	18	62	ŋ	60	28	12

a) Synthetic: Mean correct response = 68%.
b) Natural: Mean correct response = 55%.

While the responses to the synthetic nasal consonants do not look particularly decisive, they do compare favorably with similar measurements on natural nasal consonants (MALÉCOT). A confusion matrix for the latter are shown in Table 7.1 b. In this case the synthetic nasals are discriminated better than the natural ones! In view of the high functional load that nasals, particularly /n/, carry in connected speech (see Table 1.1, Chapter I), the low discrimination scores suggest that transitions, both from and into adjacent sounds, may be highly important to nasal perception.

7.32. Absolute Identification of Syllables

A substantial amount of research has considered the perception of isolated syllables. The effort has aimed mainly at discovering the acoustic cues important to phoneme recognition. Central to the objective is the determination of the separate contribution each acoustic variable makes to speech perception, as well as an understanding of how the contributions combine in the total percept. Much of the work points up the importance of acoustic environment upon perception; that is, the perception of a given phoneme can be strongly conditioned by its neighbors.

Among the leaders in this work has been the group at the Haskins Laboratories. Many of their experiments have used synthetic syllables generated by the pattern-playback machine. The operation of this synthesizer has been described in Chapter VI, and it is shown in Fig. 6.6. As explained in Section 6.21, the device synthesizes sound from data displayed as a conventional time-frequency-intensity spectrogram.

The nature of the experimentation is exemplified in consonant identification tests on CV syllables. The consonant used is either a voiced or voiceless stop. If it is voiceless (i.e., /p, t, k/), one of the variables that seems to enable listeners to differentiate the sounds is the position along the frequency scale of the brief burst of noise constituting the stop release. To isolate this particular cue and to determine its role in perception, schematized stop-vowel syllables such as shown in Fig. 7.9 c were synthesized (COOPER, DELATTRE, LIBERMAN, BORST and GERSTMAN). The noise burst (the small vertical ellipse in Fig. 7.9 c) was constant in bandwidth and duration, and the vowel was a two-formant vowel that was maintained steady throughout the syllable. Combinations of noise bursts and vowel formants shown in Figs. 7.9 a and b, respectively, produced the test ensemble.

The syllables were presented in isolation to listeners who were asked to judge the initial consonant either as /p, t or k/. The identifications, according to noise-burst location and vowel, are shown in Fig. 7.10. The contours indicate approximately equal response percentages, with the small contours representing the higher percentage response.

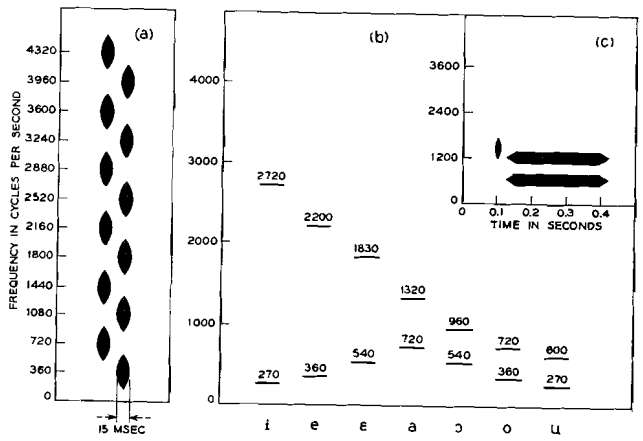


Fig. 7.9 a–c. Stimulus patterns for determining the effect of noise-burst frequency on the perception of voiceless stop consonants: (a) frequency positions of the noise burts, (b) formant frequencies of the two-formant vowels; (c) one of the synthetic consonant-vowel syllables formed by pairing a noise burst of (a) with a two-formant vowel of (b). (After COOPER, DELATTRE, LIBERMAN, BORST and GERSTMAN)

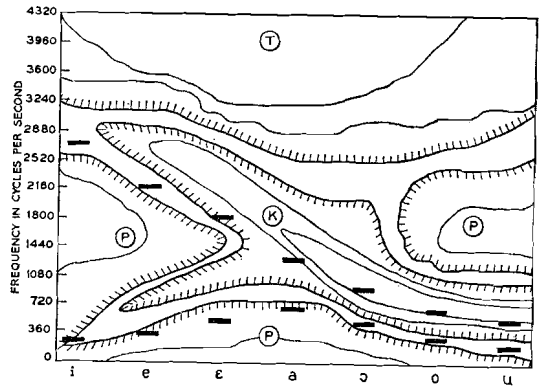


Fig. 7.10. Listener responses to the synthetic consonant-vowel syllables shown in Fig. 7.9. (After COOPER *et al.*)

For these particular syllables, the one frequency variable (namely frequency of noise burst) appears adequate to distinguish the three consonants. High frequency bursts are heard as /t/ for all vowels. For /p/ and /k/ the identification depends not only upon frequency of burst but also on its relation to the vowel. Bursts on a level with the second formant, or slightly above, are heard as /k/; otherwise they are heard as /p/. The conclusion is advanced that the perception of these stimuli

– and perhaps their spoken counterparts – requires the CV combination (that is, the syllable) as a minimal acoustic unit. Without information on the following vowel, the consonant percept may be equivocal.

A second cue important in the perception of stop-consonants is the stop-vowel formant transitions. One relevant question is how might this cue and the former one of burst position contribute singly, and how might they combine. To get some indication of the answer, the same voiceless-stop and vowel syllables were generated as before, except the noise burst was eliminated and the consonant cue was produced solely by a transition of the second formant.

The ensemble of transitions tested is shown in Fig. 7.11. The transition numbers, N, ranging from -4 to $+6$, indicate the starting frequencies of the second formant. In terms of actual cps, the starting frequencie

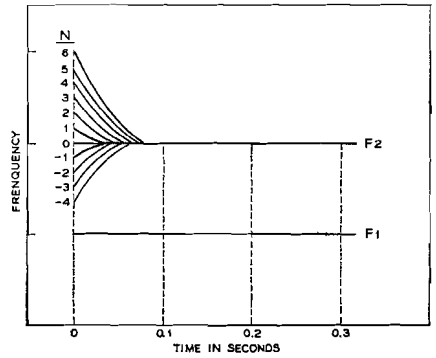


Fig. 7.11. Second-formant trajectories for testing the contribution of formant transitions to the perception of voiceless stop consonants. (After COOPER *et al.*)

are given by $[F2+N(120)]$ cps, where $F2$ is the steady-state second formant frequency of the two-formant vowels shown in Fig. 7.9[1]. The first formant was maintained constant at the values given in Fig. 7.9. The fundamental frequency of the sound was also held constant at 120 cps. The durations of the transitions were 40 msec for ± 1, and 80 msec for $+6$. For transitions in between, the durations varied linearly. The form of the transition curve is unspecified except that an effort was made to approximate the transitions seen in spectrograms of real speech. In the experience of the authors, variations in the duration

[1] An exception, apparently, was the negative $F2$ transitions of the vowels /o/ and /u/. This was $\left[F2+N\left(\frac{120}{2}\right)\right]$ (see LIEBERMAN, DELATTRE, COOPER and GERSTMAN).

of the transition and its curvature do not cause the sound to change from one stop consonant to another.

The median /p, t, k/ responses of 33 listeners, for these transitions coupled with seven different vowels, are shown in Fig. 7.12. The lengths of the plotted bars show the quantile ranges of the responses. The results indicate that the second formant transition effectively cues the /p, t, k/ discrimination.

In extending this line of investigation to other consonants, the same authors found that the second formant cues also apply to the voiced

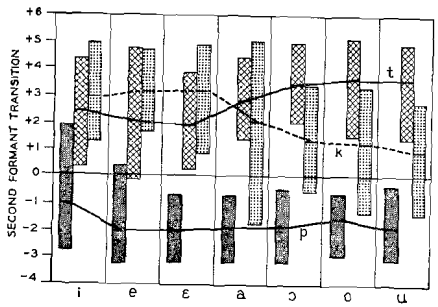


Fig. 7.12. Median responses of 33 listeners to stop consonant and vowel syllables generated by the patterns shown in Fig. 7.11. The bars show the quartile ranges. (After COOPER *et al.*)

cognates /b, d, g/. Distinctions between the voiced and unvoiced cognates are made by the first formant transition and by the voice bar. When vowel plus nasal-consonant syllables are generated in a similar manner, but with the formant transitions at the ends of the vowels and with an added, constant nasal resonance, the second formant transitions that serve to distinguish /p, t, k/ and /b, d, g/ also serve to distinguish /m, n, ŋ/ (LIBERMAN, DELATTRE, COOPER and GERSTMAN).

Returning to the syllables composed of voiceless stop and vowel, several remarks can be made. The two sets of results show the individual contributions of the noise burst in the stop release and the formant transition in the following vowel. The results do not, however, suggest how these cues combine and how they may relate to each other. One might expect that identification would be improved by the combined burst and transition cues, and that they might complement each other; when one is weak, the other might be strong. In some syllables both cues may not be sufficient, and a still different factor, such as third formant transition, may be vital to the discrimination.

The dependence of consonant perception upon the following vowel suggests to the authors that listeners perceive speech in acoustic units

of syllable length or perhaps half-syllable length[1]. A one-to-one correspondence between sound and phoneme is not found, and the phoneme may not exist in the speech wave in a free form. Clearly, one should not expect to find absolute acoustic invariants for the individual phoneme.

The experiments of the preceding discussion concerned sounds generated from abstracted spectrograms and by a particular synthesizer. Similar experiments have aimed to determine the perceptual adequacy of other synthesizers and to examine the influence of still different acoustic cues upon recognition. One of these has treated the synthesis of isolated fricatives and fricative-vowel syllables (HEINZ and STEVENS). Fricative consonants were generated by filtering noise with a single pole-zero electrical circuit. The frequency of the zero was always maintained an octave below that of the pole. The object was to determine whether such an idealized spectral representation can elicit fricative responses, and further, to establish the ranges of pole-zero locations associated with the particular responses. (Recall from Chapter III that the mode pattern of fricatives usually involves a number of poles and zeros. Recall, too, that the discussion in Section 7.27 suggests that many of the modes may not be perceptually significant.)

In one test, fricative consonants were generated and tested in isolation. A range of tuning and bandwidth was explored for the pole and zero. Identifications were made from an ensemble of five phonemes; namely, /ʃ, ç, s, θ, and f/. The synthetic sounds were 200 msec in duration. The results show that different resonant bandwidths, ranging in Q from about 5 to 10, produce no significant changes in the fricative responses. Changes in tuning of the resonance, however, produce important differences in response. The effect is illustrated by the percentage response *vs* resonant frequency plotted in Fig. 7.13. The /f/ and /θ/ responses are combined.

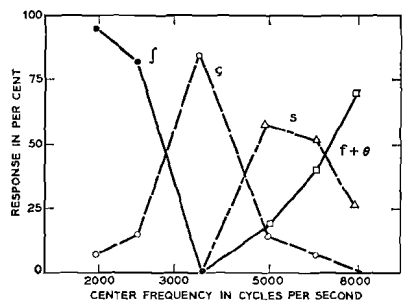


Fig. 7.13. Listener responses in absolute identification of synthetic fricatives produced by a pole-zero filtering of noise. The frequency of the pole is indicated on the abscissa, and the frequency of the zero is approximately one octave lower. (After HEINZ and STEVENS)

[1] This point, and other views on it, will be discussed further in Section 7.5.

Using the same synthetic fricatives, consonant-vowel syllables were synthesized with a terminal-analog synthesizer. The vowel used was always /ɑ/, and the syllable synthesized is illustrated by the schematic spectrogram in the upper part of Fig. 7.14. The timing sequence of control functions for the terminal-analog synthesizer is shown by the lower curves in Fig. 7.14. The first two curves show the build-up and decay characteristics of the noise (voiceless) and buzz (voiced) excitation. The third curve shows the timing of the formant transitions. The $F1$ vowel transition always started from 200 cps. The initial $F2$ value was either 900, 1700 or 2400 cps. Fricative resonances of 2500, 3500, 5000, 6500 and 8000 cps were used. Listeners were required to identify the initial consonant as /f, θ, s, or ʃ/.

The consonant judgments—as functions of the fricative resonance frequency and second-formant transition—are plotted in Fig. 7.15. The results for two ratios of consonant-to-vowel intensity are shown,

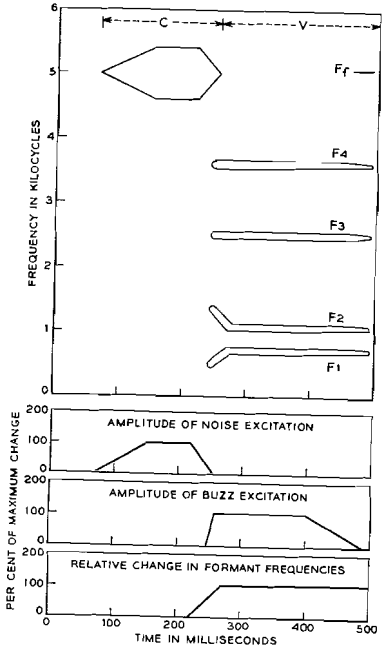


Fig. 7.14. Abstracted spectrogram showing the synthesis of a syllable with fricative consonant and vowel. The single fricative resonance is F_f. The four-formant vowel is an approximation of /ɑ/. The lower three curves represent the temporal variation of the excitation and formant frequencies in the syllable. (After HEINZ and STEVENS)

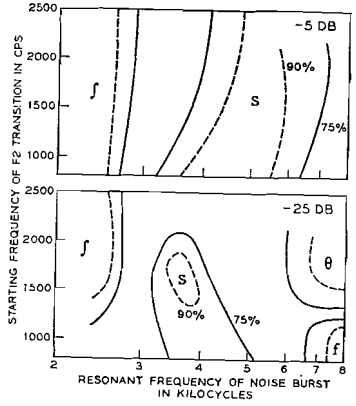


Fig. 7.15. Absolute identifications of the initial consonant in the synthetic syllable schematized in Fig. 7.14. Two response contours are shown corresponding to 90 and 75% identification. Two consonant-to-vowel intensities (−5 and −25 db) are shown. (After HEINZ and STEVENS)

namely −5 db and −25 db. Two response contours are also shown. Inside the dashed lines the indicated fricative is responded in more than 90% of the presentations. Inside the solid lines the response is greater than 75%. The two consonant-to-vowel intensities dramatize the importance of relative level in the perception of /θ/ and /f/, and to a lesser extent, /s/. The responses also suggest that the fricative /f/ is distinguished from /θ/ largely on the basis of the $F2$ transition in the vowel. Contrariwise, the formant transition does not have much influence upon the /s/ and /ʃ/ discrimination, this being determined more by the frequency of the fricative resonance. Another study, closely related in form and philosophy to the present one, has been carried out for Japanese fricatives (NAKATA, 1960).

In much the same vein, other experiments have studied formant transitions with a transmission-line analog (STEVENS and HOUSE, 1956). The results show that low $F2$ loci (1000 cps or less) are generally associated with bilabial or labio-dental articulatory configurations. On the other hand, $F2$ loci in the middle frequency range (1500 to 2000 cps) are associated with alveolar configurations, and $F2$ loci above 2000 cps are associated with palatal configurations.

A still different approach to synthesis and perception is exemplified by the generation of connected speech from individual, spectrally-constant synthetic segments (COHEN and 'T HART). The segments are of phoneme length and are time-gated with prescribed build-up, decay and duration. From these results the suggestion is advanced that proper dimensioning of the time parameter makes it possible to neglect a number of details of formant information usually considered to be of paramount importance. It seems reasonably clear, however, that the ear accomplishes a short-time spectral analysis (see Chapter V) and that it appreciates continuous variations both in frequency and intensity. The "time parameter" view implies a trading relation of a sort between spectral information and temporal detail. Such a trade may in fact exist, but the extent to which it can be exploited may be limited. It would appear unlikely that high-quality, high-intelligibility speech could be consistently synthesized without taking account of mode transitions within phoneme-length segments.

7.33. Effects of Learning and Linguistic Association in Absolute Identification of Speech-Like Signals

It was suggested earlier that at least two limitations exist in applying classical psychophysical data to speech recognition. First, the classical measures are generally restricted to differential discriminations. Second, they are usually made along only one dimension of the stimulus. Speech,

however, appears to be a multidimensional stimulus. Its perceptual units, whatever they might be—and they probably vary according to the detection task—are presumably perceived absolutely. At least one experiment has attempted to measure the effects of learning and linguistic association in absolute discriminations. The tests treated several dimensions of complex, speech-like sounds (HOUSE, STEVENS, SANDEL and ARNOLD).

Four different groups of stimuli (*A*, *B*, *C* and *D*), varying in their similarity to speech, were used. The stimuli of each group were further divided into subgroups. The signals of each subgroup were coded in a given number of dimensions. Each member of the subgroup was designed to convey three bits of information per presentation. The signals of the *A* group, for example, were produced by filtering random noise with a simple resonant circuit. They could be coded along time, frequency and intensity dimensions. Stimuli in subgroup *A*1 were coded unidimensionally in terms of 8 frequency positions of the noise resonance. The center frequency of the resonance varied from 500 to 5000 cps, and its corresponding bandwidth varied from 300 to 3120 cps. One intensity (namely, a reference intensity) and one duration (300 msec) were used. In contrast, stimuli of subgroup *A*7 were coded in terms of two frequency positions of the noise (820 or 3070 cps), two intensity values (± 8 db *re* *A*1), and two durations (150 or 450 msec). The subgroups *A*2 through *A*6 utilized different combinations of dimensions and quantizations between these extremes.

The *B* stimuli were also rudimentary signals but with slightly more speech-like properties. They had temporal and spectral properties roughly analogous to vowel-consonant syllables. The vowel element was produced by exciting a single resonant circuit with 125 pps pulses. The center frequency of the resonator was 300 cps and its bandwidth was 60 cps. The consonant portion was produced by exciting a simple resonant circuit with white noise. The coded dimensions of the *B* signals were center frequency and bandwidth of the noise portion (center frequencies 500 to 5000 cps, bandwidths 100 to 1000 cps); intensity of noise (± 14 db); and duration of the silent interval (gap) between the vowel and consonant (10 to 180 msec). The total duration was always 350 msec. Like the *A* group, set *B*1 was a one-dimensional coding and had eight frequency values, one intensity and one duration. Set *B*7 was a three-dimensional coding and had two frequencies, two intensities and two gap durations.

The *C* group was constructed to be still more similar to speech. It incorporated many of the characteristics of acceptable synthetic speech samples. Like *B*, the *C* stimuli were vowel-consonant syllables, but the vowel was generated from four resonators whose center frequencies

were fixed at 500, 1500, 2500, and 3350 cps. Their bandwidths were approximately those of spoken vowels. The first formant was given a falling transition to the time gap, in analogy to the vowel-to-stop consonant transition. The consonant portion was generated by a single pole-zero filtering of noise, similar to the circuit described in the preceding section for producing fricative consonants (HEINZ and STEVENS). Voiced excitation during the vowel was inflected from 120 to 150 pps. The stimulus dimensions and the varied parameters were similar to those of the *B* signals. In set *C*1, the consonant resonance varied from 500 to 5000 in eight steps. The vowel duration was 250 msec, the gap 50 msec, and the consonant 100 msec. (Total duration was always 400 msec.) In set *C*7, the consonant dimensions of resonance, intensity and gap were all binary.

The *D* stimuli were real, monosyllabic speech utterances produced by one speaker. Only a single, three-dimensional subgroup was used. The eight syllables were composed of two vowels, /I/ and /*A*/, and four consonants /f, s, p, t/. Four of the eight syllables were monosyllabic English words, and four were nonsense syllables.

In the tests the stimuli were presented singly in isolation. Listeners were required to associate each with one of eight unlabelled buttons on a response panel. After the subject made his selection, one of eight lights on the panel flashed, indicating the correct button with which to associate the stimulus. The next sound was then presented. There was no speed requirement.

The results show how the median probability of correct identification increases with learning. Identification data from twelve listeners for the unidimensional, frequency-coded stimuli are shown in Fig. 7.16. Each test block involved the randomized presentation of sixteen items from a given (8-component) stimulus ensemble. The responses to the tri-dimensional stimuli are given in Fig. 7.17.

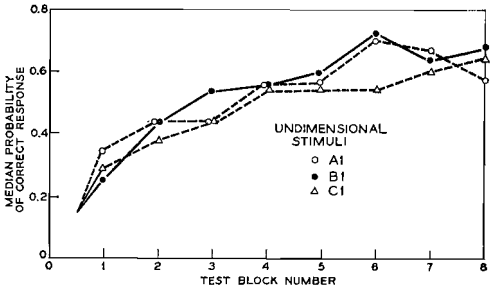


Fig. 7.16. Median probability of correct response for frequency-coded, one-dimensional stimuli. (After HOUSE, STEVENS, SANDEL and ARNOLD)

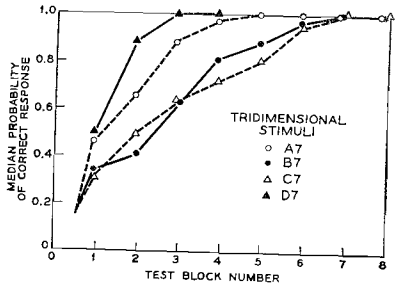


Fig. 7.17. Median probability of correct response for time-frequency-intensity coded three-dimensional stimuli. (After HOUSE, STEVENS, SANDEL and ARNOLD)

The two sets of results show that learning is more rapid for the tri-dimensional stimuli than for the one-dimensional items. Of the tri-dimensional signals, real speech (*D*7) is learned the fastest. The least speech-like artificial signal (*A*7) is learned the next fastest. The results suggest two conclusions. First, performance during learning is better when the stimuli are coded in several physical dimensions than when they lie along a unidimensional continuum. Second, as the physical characteristics of the stimuli are made more similar to speech, there is a deterioration of performance, except for stimuli that are actual speech signals!

The explanation advanced for this latter and somewhat surprising result is that neither the *A*, *B*, nor *C* stimulus ensembles were sufficiently like speech to elicit a linguistic association. Hence, they had to be identified in a manner different from speech. Real speech sounds, however, are categorized with great facility by listeners, and presumably the subjects made use of linguistic categories in discriminating the *D* stimuli. The *A*, *B*, and *C* signals, lacking linguistic association, were probably identified in terms of what may be more "natural" basic dimensions in perception, namely, loudness, pitch and duration. Discrimination of these fundamental dimensions might be expected to be more clear cut for the *A* stimuli. The *B* and *C* signals apparently do not order well along these dimensions because of the fixed initial vowel segment.

The results are therefore interpreted to argue against the existence of a speech-like continuum. Although the signals may bear more or less resemblance to speech from a physical point of view, the subjective responses exhibit a sharp dichotomy. Either the sounds are associated with linguistic entities or they are not. In the present experiment presumably none of the synthetic sounds were associated with linguistic quantities. Within a linguistic frame, the tendency is to categorize a

signal according to dimensions established by the language structure. Perception of the signal as a linguistic unit probably depends strongly upon nonperipheral processes. Small details of the signal, routinely preserved at the periphery of the ear, may not be of primary importance. For nonlinguistic signals, on the other hand, the tendency is to order them along what seem to be natural psychological dimensions. Their discrimination probably requires less central processing than does the perception of speech.

7.34. Influence of Linguistic Association Upon Differential Discriminability

A listener's linguistic learning and experience provide an acute ability to categorize speech signals. In the experiment of the preceding section, listeners presumably resorted to linguistic associations for the *D*7 stimuli. They apparently did not for the other stimuli, either because the signals were not sufficiently speech-like, or because the listener's attention was not drawn to such an association by the instructions given him.

The results therefore raise a further question. Assuming that a linguistic association is made, is its effect reflected in the differential discriminations a listener can make? In other words, can the learning and discriminability acquired in linguistic experience carry over into a more classical differential comparison. At least one experiment suggests that it can (LIBERMAN, HARRIS, HOFFMAN and GRIFFITH). The objective was to demonstrate that the differential discriminability of formant motion in a synthetic speech syllable is more acute when the change traverses a phoneme boundary.

Consonant-vowel syllables were synthesized with the pattern playback device described in Section 6.21, Chapter VI. Two formants were used and the vowel was always /e/ ($F1=360$, $F2=2160$ cps). The consonants were various two-formant transitions spanning the known approximations to /b, d, and g/. The set of synthetic syllables used is shown in Fig. 7.18. The positive first-formant transition is the same in all the syllables and is a necessary cue to voicing. The second formant

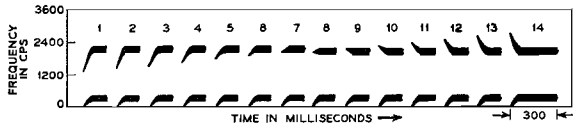


Fig. 7.18. Synthetic two-formant syllables with formant transitions spanning the ranges for the voiced consonants /b, d, g/. The vowel is the same for each syllable and is representative of /e/. (After LIBERMAN, HARRIS, HOFFMAN and GRIFFITH)

transitions range from highly negative to highly positive. The duration is the same for all syllables, namely 300 msec.

Two tests were made. In one, the stimuli were presented singly for absolute judgment of the consonant. The allowed response categories were /b, d, and g/. In the second, an *ABX* presentation was made. Stimuli *A* and *B* were different syllables from Fig. 7.18. They were separated by either one, two or three successive steps shown in Fig. 7.18. Sound *X* was identical to either *A* or *B*. On the basis of any cues they chose to use, listeners judged whether *X* was most like *A* or *B*. The second test therefore gave a measure of relative discriminability at each step on the continuum described by the stimuli in Fig. 7.18.

The absolute identification results of the best subject in the experiment are shown in Fig. 7.19. This same subject's responses in the *ABX* test, when the step size between *A* and *B* is two (that is, the *B* stimulus number is *A* plus two in Fig. 7.18), are given in Fig. 7.20. Comparison of the data shows a clear diminution of differential discriminability of formant transition for the stimuli contained *within* the /b/ and /d/ response ranges. A corresponding drop for the /g/ range apparently is not obtained. The other subjects in the experiment did not give data with maxima and minima so well defined, but the indications are that somewhat similar variations exist.

A rough approximation of differential discriminability can be made on the assumption that listeners can discriminate only so well as they can identify. This assumption tends to predict the relative variations in discriminability, but it underestimates the absolute level of discriminability. The difference may represent a so-called margin of true dis-

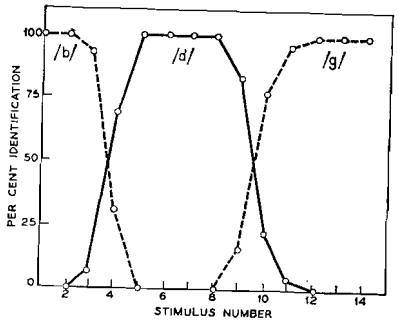


Fig. 7.19

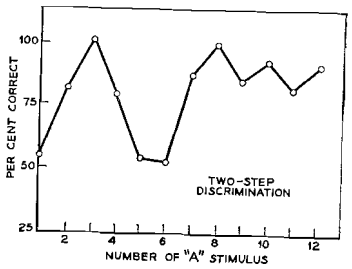


Fig. 7.20

Fig. 7.19. Absolute consonant identifications of one listener for the stimuli of Fig. 7.18. (After LIBERMAN *et al.*)

Fig. 7.20. *ABX* responses of the listener whose absolute responses are shown in Fig. 7.19. The step size between *A* and *B* stimuli was two positions in the stimulus set of Fig. 7.18. (After LIBERMAN *et al.*)

crimination, that is, the ability of listeners to distinguish speech sounds not solely on the basis of phoneme labels, but also more directly by acoustic differences.

The suggestion is advanced that the inflection points in discrimination are not innately built into the human. Different languages have phoneme boundaries in different places. The case for acquired discriminability would of course be strengthened by demonstrating that native speakers of other languages exhibit maxima of differential sensitivity placed along the continuum in accordance with their languages. The crucial factor in the present experiment is the extent to which linguistic associations are elicited by the stimuli[1]. Lacking the ability to categorize, the differential discriminability might be expected to be monotonic along the stimulus continuum.

To inquire into this last point, a similar experiment was conducted on synthetic vowel sounds (LIBERMAN, COOPER, HARRIS, and MACNEILAGE). No increase in discrimination was found at the phoneme boundaries. In addition, the differential discriminability lay considerably above that predicted simply on the basis that listeners can discriminate only so well as they can identify. (In other words, listeners can discriminate many within-phoneme differences.) The conclusion is that the perception of vowels tends to be continuous and is not as categorized as, for example, the stop consonants. A further experiment with two other phonemic distinctions, namely vowel length and tone in Thai, also failed to show sharpening at the phoneme boundary (LIBERMAN, COOPER, HARRIS, and MACNEILAGE).

7.4. Effects of Context and Vocabulary Upon Speech Perception

The precision with which listeners identify speech elements is intimately related to the size of the vocabulary and to the sequential or contextual constraints that exist in the message. The percent correct response is higher the more predictable the message, either by virtue of higher probability of occurrence or owing to the conditional probabilities associated with the linguistic and contextual structure. This influence is apparent in intelligibility scores for various types of spoken material. Fig. 7.21 illustrates the effect in an experiment where speech was masked by varying amounts of noise (MILLER, HEISE, and LICHTEN).

Three different types of test material were used. Articulation tests were made with the same subjects and experimental apparatus. One set of material was the spoken digits zero to nine. Another was complete

[1] The question is made more pointed, perhaps, by the results of the previous section where apparently no linguistic association was made with synthetic syllables.

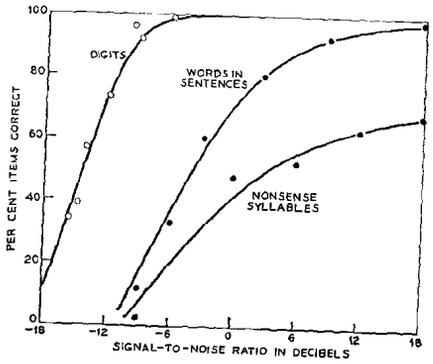


Fig. 7.21. Intelligibility scores for different types of spoken material as a function of signal-to-noise ratio. (After MILLER, HEISE and LICHTEN)

sentences read and scored for the major words. A third was nonsense syllables which were pronounced and recorded using an abbreviated phonetic notation. As Fig. 7.21 shows, the signal-to-noise ratios necessary to produce 50 percent correct response are approximately -14 db for the digits, -4 db for the words in sentences, and $+3$ db for nonsense syllables. The discriminations among a small number of possibilities are obviously better than among a large number. The sequential constraints present in the sentences apparently result in higher intelligibility scores than for the nonsense material.

The effect of vocabulary size was examined in further detail. The same type of articulation tests were performed on monosyllabic word sets numbering 2, 4, 8, 16, 32, 256, or an unspecified number. For the restricted vocabularies, the listeners were informed of the alternatives. The results of the intelligibility tests are shown in Fig. 7.22. The results show clearly that as vocabulary size increases, the signal-to-noise ratio necessary to maintain a given level of performance also increases.

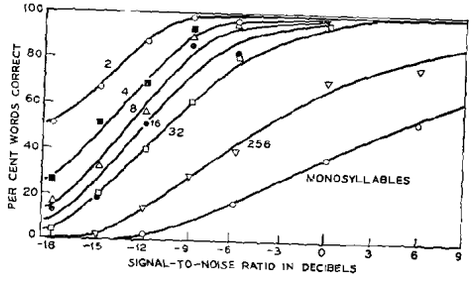


Fig. 7.22. Effects of vocabulary size upon the intelligibility of monosyllabic words. (After MILLER, HEISE and LICHTEN)

Semantic and syntactical constraints also influence the predictability of a speech utterance and hence its intelligibility. The grammatical rules of a given language prescribe allowable sequences of words. Semantic factors impose constraints upon those words which can be associated to form a meaningful unit. Experiments have demonstrated that the intelligibility of words is substantially higher in grammatically-correct, meaningful sentences than when the same words are presented randomly in isolation (MILLER, HEISE, and LICHTEN). The sentence context reduces the number of alternative words among which a listener must decide, and the improvement in intelligibility is due, at least partially, to this reduction.

Reduction in the number of alternatives, however, is not the sole factor. Experiments have compared the intelligibility of words in grammatically-correct, meaningful sentences to the intelligibility in nongrammatical, pseudo-sentences (G. A. MILLER, 1962). The pseudo-sentences were constructed so that the number of word alternatives was exactly the same as for the grammatical sentences. In the grammatical structures a listener apparently accomplishes perception in terms of phrases, or longer elements. He may delay decisions about words, rather than make them about each word as it occurs. The nongrammatical structures, on the other hand, cannot be processed this way. They must be perceived in terms of shorter temporal elements.

A somewhat different emphasis can be placed on context from the standpoint of acoustic environment and reference. Many perceptual evaluations seem to be made by a relative rather than absolute assessment of physical properties. That is, the physical surround establishes a frame of reference for the decoding operation. A simple example might be the pitch inflection of an utterance. The relative change, or pattern of inflection, is probably more significant perceptually than the absolute number of cycles per second.

Such acoustic "referencing" has been demonstrated in synthetic speech. It can be present to the extent that identification of a given monosyllabic word is strongly influenced by the time-frequency-intensity frame within which it is placed (LADEFOGED and BROADBENT). For example, a given synthetic vowel was produced as the central element of the synthetic word /b–t/. This word was used in synthetic sentences having different relative patterns of formant frequencies. Depending upon the acoustic reference established by the formant patterns in the rest of the sentence, the physically same synthetic word was variously identified as *bit*, *bet* or *bat*.

7.5. The Perceptual Units of Speech

The data in the preceding discussions suggest that speech perception is an adaptive process. It is a process in which the detection procedure

probably is tailored to fit the signal and the listening task. If the listener is able to impose a linguistic organization upon the sounds, he may use information that is temporally dispersed to arrive at a decision about a given sound element. If such an association is not made, the decision tends to be made more upon the acoustic factors of the moment and in comparison to whatever standard is available.

The suggestion that a listener uses temporally spread information raises the question as to the size of the temporal "chunks" in which speech is perceived. Very probably the size of the perceptual element varies with the discrimination task, and the listener adjusts his processing rate to suit different types of speech information. He may, for example, attend to prosodic information while phonemic information is momentarily predictable. For nonspeech or nonlinguistically associated discriminations, the perceptual processing may be substantially different. In either case, however, the information must funnel through the same sensory transducer. As mentioned earlier, differential discriminations of "classical" psychoacoustic signals probably reflect the fundamental limitations of the transducer and the peripheral processing, whereas linguistically-connected discriminations probably reflect the storage and processing characteristics of the central mechanism.

Speech recognition presumably requires that sound elements be identified in absolute terms. For some sounds, however, distinctiveness is not so much an acoustic, or even articulatory factor, but a consequence of linguistic experience. A distinctiveness, which may be salient in connected speech, may be diminished or altogether lost in isolation. A case in point concerns the nasal consonants. These sounds carry a heavy functional load in connected speech (see Table 1.1, Chapter I), but are poorly identified in isolation (see Table 7.1, Section 7.31).

A number of studies have aimed at determining the units in which perception occurs. For the most part the experiments arrive at disparate results, probably owing to the large differences in perceptual tasks and to the fact that there may be no single answer to the question. Perhaps exemplifying one extreme in perception is the task of speech "shadowing" (CHISTOVICH, 1962). This approach aims to resolve whether, upon hearing the beginning of a speech sound, a listener immediately begins to make some preliminary decisions and corrects them as more information becomes available, or whether he stores long portions of data before interpreting them. The question was examined in two ways. First, the latency was measured for the articulatory movements of a listener who was repeating as rapidly as possible ("shadowing") the speech syllables he heard over earphones. The syllables were either vowel-consonant-vowel or consonant-vowel. Second, the latency was measured for a written transcription of the consonant sounds in the syllables heard.

The results showed that in the vocal shadowing, the consonant latencies were on the order of 100 to 120 msec for the *VCV* syllables, and on the order of 150 to 200 msec for the *CV*'s. In the *VCV* syllables the subject apparently anticipates the *C* before it is completely articulated, perhaps getting a good deal of information from the formant transitions in the initial *V*. He is often wrong initially, but generally corrects himself (on a running basis) by the end of the *C*. Because the subject reacts before he perceives the whole consonant—and even makes responses that are not possible in his language—the interpretation is advanced that the subject makes a number of simple decisions about the articulatory origin of the acoustic event (that is, whether the origin is dental, voiced, voiceless, nasal, etc.). The decisions are corrected as the sound proceeds, and a set of features are finally accumulated to form the phoneme. It is therefore suggested that shadowing is "phoneme creation" from simple decisions about articulatory parameters.

The latencies for the written mode of response were found to be very nearly the same as the latencies to the ends of the *C*'s in shadowing (that is, the interval between ends of the original and the shadowed *C*'s). The conclusion is therefore put forward that consonant writing is closely related to consonant shadowing.

It is difficult to say precisely how perception under these conditions relates to perception of running speech. The results may be strictly interpretable only within the frame of the task. If the task is made different, the measures are likely to indicate a different duration for the "unit". Another experiment perhaps illustrates the opposite extreme in evaluating the unit. It suggests that listeners are not only aware of large parts of an utterance at any moment, but actually may find it difficult to consider speech in terms of small segments, even when asked to make an effort to do so (LADEFOGED).

The spoken word "dot" was superimposed on the recording of a complete sentence. Listeners were asked to note and report the precise moment in the sentence when the superimposed word commenced. The judgments were generally inaccurate, but it was not uncommon for subjects to report that the superimposed item occurred two or three words earlier in the sentence than was actually the case.

This behavior suggests that the mechanisms and times for processing on-going contextual information may be considerably different from those for isolated stimuli, even though the latter are speech sounds. It also suggests that continuous speech produces complex temporal patterns that are preceived as a whole. Items such as syllables, words, phrases, and sometimes even sentences, may therefore have a perceptual unity. In such an event, efforts to explain perception in terms of se-

quential identification of smaller segments would not be successful. As a consequence, attempts to build machines that recognize speech in terms of brief acoustic units may be of little or no profit.

It was suggested earlier (see Section 7.33) that "natural" auditory dimensions apparently include subjective attributes such as pitch, loudness, and temporal pattern, and that these dimensions appear useful in discriminating nonlinguistically associated sounds. These same dimensions may of course apply to continuous speech signals, but they may be assessed in different ways—perhaps in ways that are related to production. For example, there is some evidence that the loudness of speech is assessed more in terms of the respiratory effort required to produce the necessary subglottal pressure than it is in terms similar to, say, the loudness scale for sine waves (LADEFOGED). If the "motor theory" of speech perception has validity, a listener may evaluate a speech signal in terms of the motor activity that produced it, as well as in terms of other acoustic factors not directly under motor control.

Many theorists in speech perception appeal to a link between production and perception. How tight this link is, is not known. If it is close, perception could conceivably occur in terms of "articulatory" segments rather than acoustic segments. In producing speech, the human has at least three kinds of feedback: auditory, tactile and proprioceptive. Blocking of one or more of these channels apparently causes some of its functions to be assumed—but generally less well—by one of the other channels. Speech attributes such as vowel quality, nasality and pitch seem highly dependent upon auditory feedback, while features such as lip and tongue movements in consonant articulation seem more dependent upon tactile and proprioceptive channels. If perception is linked to these processes, some speech properties might be identified by reference to acoustic factors, and others by reference to articulatory activity.

7.51. Models of Speech Perception

Much progress remains to be made in understanding and in modeling the mechanism of human speech perception. Not least is the problem of quantifying behavior in response to speech signals. Appeal to the mechanism of speech production is sometimes made on the basis that perceptual factors, at some level, must correspond to those necessary to speak the same message. This "motor theory of speech perception" has been the focus of considerable speculation and not little controversy (LIBERMAN *et al.*). If truly invoked by humans—which has not been shown—it has the advantage that motor commands to the vocal mechanism are more amenable to psychological study than are, say, electrical representations of speech signals in the human contex. Further, acoustic

and linguistic correlates of the motor commands are more accessable for study.

At least one view (BONDARKO *et al.*) has maintained that the development of a model of human speech perception is the same problem as the development of an automatic speech recognizer, and further, that present knowledge embraces only the most rudimentary aspects of such a model. The proposal for such a model involves the hierarchial structure shown in Fig. 7.23. The model is envisioned as a chain of transformations in which each stage acts as an information filter to

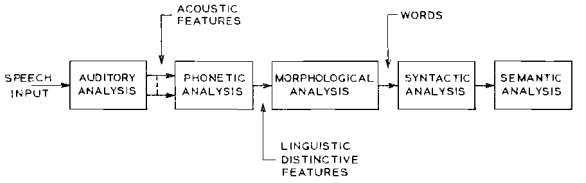


Fig. 7.23. Block diagram model of stages in speech perception.
(After BONDARKO, ZAGORUYKO, KOZHEVNIKOV, MOLCHANOV and CHISTOVICH)

reduce the dimensionality of the signal. For example, the first three blocks transform an acoustic signal into a succession of words where each word is described by a set of lexical and grammatical features and by prosodic characteristics. Syntax and finally semantic analysis complete the transformations necessary for message understanding. The natures of the transformations, if in fact they exist in identifiable and separable forms, are not known. Perceptual experiments do, however, suggest certain characteristics of the first two stages.

The peripheral auditory analysis made by the human cochlea is such that features of the short-time spectrum of the input signal are preserved. This analysis preserves temporal detail relevant to changes in spectral distribution, periodicity (or non-periodicity) and intensity. That this is true can be shown by psychoacoustic experiments on perception of changes in pitch, formants or intensity of speech and speech-like sounds. That this information is reduced in "dimensionality" for later processing is supported by experiments which show that consonant perception is influenced only by the direction and rate of change of formant transitions, and not by absolute values of their "loci" or initial frequencies. Similar perceptions of the direction and rate of change of fundamental frequency, or pitch, influence nasal-non-nasal discriminations in labial consonants (CHISTOVICH).

The reduction of dimensionality performed in the phonetic analysis is likely to be one of feature analysis rather than one of comparison

to a stored reference pattern. This view is supported by data on syllable recognition where features such as manner of production may be perceived correctly while, say, place of production is perceived incorrectly. Similarly, prosodic features may be perceived without discrimination of phonetic factors. Experiments on mimicking and shadowing (CHISTOVICH, KOZHEVNIKOV, and ALYAKRINSKII) are consistent with this in that some phonematic features can be recognized and produced even before a listener hears a whole syllable. This type of feature analysis also argues that the input to the phonemic analysis block of Fig. 7.23 may already be organized for parallel, multichannel processing.

Exactly what duration of signal may be subjected to such analysis in not clear, but data on short-term auditory memory provides some insight. In recall experiments with speech (MILLER; NEVEL'SKII) a sequence of three vowels or three tones is recalled as a sequence of decisions regarding the stimuli and not as a sequence of acoustic descriptions (CHISTOVICH, KLAAS, ALEKIN). The phonemic analysis must therefore work with speech segments shorter than average word length. Furthermore, experiments show that a man cannot remember sequences of nonsense syllables longer than 7 to 10 syllables (MILLER; CHISTOVICH, KOZHEVNIKOV, and ALYAKRINSKII). This fact bears on the size of the short-time storage and characterizes the "time window" through which the message is "seen" by the morphological analysis stage.

On the other side it is clear that a listener does not make separate decisions about every phoneme in running speech. The units with which he operates likely correspond to words, or to even longer segments. Information handed from the morphological analysis to the syntactic and semantic analysis can, consequently, be reduced in dimensionality to this extent. Auditory segments need not coincide with phonemes—i.e., each segment need not contain information about one and only one phoneme and the number of segments need not equal the number of phonemes.

Experiments on recall show that a listener remembers phonemes as a set of features (WICKELGREN; GALUNOV). Therefore, the phonemic information at the output of the phonetic analysis block should be represented by abstract, distinctive features. Several different acoustic (or auditory) features may contain information about one and the same distinctive feature.

7.6. Subjective Evaluation of Transmission Systems

7.61. Articulation Tests

A conventional technique for evaluating a communication facility is to determine the intelligibility of the speech it transmits. This is customarily done by counting the number of discrete speech units correctly recognized by a listener. Typically, a speaker reads a list of syllables, words, or sentences to a group of listeners. The percentage of items recorded correctly is taken as the articulation score. By choosing test material representative of the sound statistics of a language, a realistic test can be made of the transmission system. The development of the so-called phonetically-balanced (PB) test words has this objective (EGAN). The techniques for administering and scoring articulation tests have been described in many places, and there is little need to repeat the procedures here (see for example, BERANEK, 1954; HARRIS, ed.; RICHARDSON, ed.).

An articulation score is not an absolute quantity. It is a function of parameters such as test material, personnel, training, and test procedure. It consequently should be treated as a relative measure. Usually the significant information is a difference between scores obtained with the same material, procedures and personnel. Syllable and word items can be scored in terms of the correctness of their written response. Sentences can be scored either in terms of their meaning conveyed, or in terms of key words in the sentence. Contextual constraints usually make the scores for sentences higher than those for isolated words. One relation that has been worked out between word articulation and sentence intelligibility (in terms of meaning conveyed) is shown in Fig. 7.24 (EGAN).

Articulation tests are typically done without speed requirements, and the stimulus presentation rates are favorable for careful consideration of each item. More realistic articulation tests—so far as the informational capacity of a transmission system is concerned—should

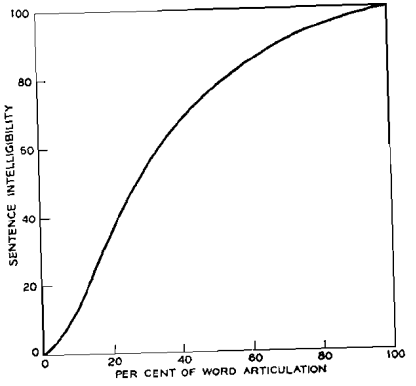


Fig. 7.24. A relation between word articulation score and sentence intelligibility. Sentences are scored for meaning conveyed. (After EGAN)

include time limitations. Some research into the design of such tests has been initiated (D'EUSTACHIO). The philosophy of adding stress to the communication task is that "fragile" systems will fail before more robust systems with perhaps valuable redundancy. Time limitation is but one way stress can be introduced. Additional mental activities, such as required with simultaneous motor or visual tasks, also load the listener. The aim is to control the sensitivity of the test by varying the subjective load (NAKATANI; MONCUR and DIRKS).

7.62. Quality Tests

In the conventional articulation test, a listener is usually required to respond with the written equivalent of the speech he hears. The quality or naturalness of the signal is not specifically evaluated. Methods for quantitatively rating speech quality have not been well established, mainly because the physical correlates of quality are poorly understood. Various rating-scales and rank-order methods have been examined (EGAN). However, generally applicable techniques for uniquely relating speech quality and acoustic factors are not presently available.

One proposal has suggested that speaker recognition is an important and measurable aspect of naturalness (OCHIAI and KATO; OCHIAI, 1958). Results along these lines suggest that spectral distortions of a speech signal affect the accuracy of speaker identification much differently from the way they affect phoneme identification. Another proposal has been to consider voice quality as the "spectral remainder" after inverse filtering a prescribed number of formants out of the signal (FUJIMURA). A large contribution to what remains is then attributed to the source of vocal excitation.

Perhaps one of the most promising tools for assessing speech quality is Multi-dimensional Scaling (SHEPARD; KRUSKAL; CARROLL). In this technique, non-metric data, corresponding to subjective judgments on a signal, are analyzed to reveal preferred rankings, and to show how individual subjects weight (in importance to preference) different attributes of the stimulus.

The technique assumes that observers use a common set of subjective factors (or coordinates) on which to base their judgements. The analysis indicates the number of such factors needed to account for prescribed amounts of variance in the subjective judgments. It does not, however, identify the physical correlates of the factors. This interpretation is a human one, and must rest upon knowledge of the physical properties of the stimuli.

The method is applicable to judgments made in close comparison (say, similarity or difference judgments on stimulus pairs) and to judgments made on an absolute basis (say, absolute assignment of quality

ratings). Numerous variations of the method exist. An explanation of all would fill a book itself. The most expedient vehicle to illustrate the nature of the method is a specific example.

In one application, multidimensional scaling was used to assess the acceptability of amplitude-modulated, periodic pulses as an electronic telephone ringing signal (BRICKER and FLANAGAN). Physical variables were pulse repetition frequency (f_0), harmonic content (c), modulation frequency (f_m) and modulation duty-factor (df)[1]. Listeners heard single presentations of each signal condition and assigned an absolute numerical rating chosen from an unbounded range of positive and negative integers. Positive ratings were assigned to signals that were liked and negative to those disliked. The assigned ratings of each subject were converted to standard scores having zero mean and unity standard deviation.

The normalized judgments of n subjects on m different signal conditions produce an $n \times m$ data matrix $\boldsymbol{S}$. The multidimensional procedure factors this data matrix into an $n \times r$ matrix of subject vectors and an $r \times m$ matrix of stimulus coordinates in r-dimensional space. The product of the subject and stimulus matrices is an $n \times m$ matrix $\boldsymbol{S}^*$ which is, in a least-squares sense, the best approximation of rank r to the original data matrix $\boldsymbol{S}$. In particular, the r-dimensional projections of the stimuli onto each subject's vector constitute the best approximation to that subject's original data vector. The r-dimensional projections of a subject's vector onto the r orthogonal coordinates indicate the relative weights assigned to the coordinates by that subject.

The goal is to find directions in r-space along which signals are ordered in a physically interpretable manner. These directions are then related to the common perceptual attributes assumed as the basis for judgment. The relation of the subject vectors to these directions indicate the weight (or importance) of the attributes in the individual subjective ratings.

The r-dimensions are ordered according to the size of their characteristic roots, or to the proportion of the variance they account for in the original data. In the present example 40 subjects rated 81 signal conditions, and three dimensions accounted for most of the variance ($r=3$). The projections of the subject vectors onto the two most important dimensions are shown in Fig. 7.25a.

Each arrowhead is the endpoint of a unit vector in the 3-dimensional unit sphere generated by the program. The vector thus specified may be imagined as a line segment from the end point extending through the origin and an equal distance beyond; the arrow points in the direction of higher rating by that subject. The relative weights given to each

[1] The modulation waveform was a half-wave vectified version of $(a+\sin 2\pi f_m t)$. The constant a was used to control duty factor.

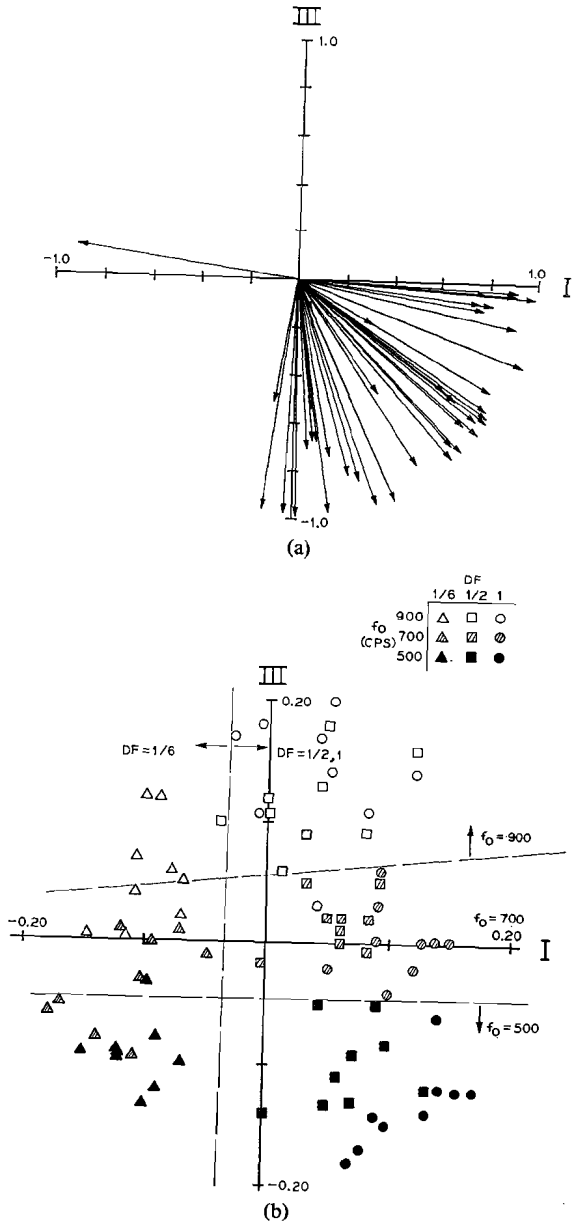


Fig. 7.25a and b. (a) Subject vectors obtained from a multi-dimensional scaling analysis projected onto the two most important perceptual dimensions I and III. The data are for a tone ringer experiment. (b) Preference judgments on 81 tone-ringer conditions, projected onto the two most important perceptual dimensions I and III. Direction of high preference is indicated by the vectors in Fig. 7.25a. (After BRICKER and FLANAGAN)

of the three dimensions by a given subject, according to the assumptions of the technique, are reflected graphically by the perpendicular projections on the three axes of that subject's endpoint. Specifically, the squares of the projected values sum to 1.0 (by definition of the unit vector) and the subject weights are quantitatively related as the squares of the projected values. Thus, a subject whose endpoint is close to the end of one axis is described by the model as weighting that dimension heavily and the other two negligibly. One subject in Fig. 7.25a is seen to assign weights particularly different from the other 39.

The 81 stimulus coordinates of the preference judgments on the 81 signal conditions are shown projected onto the same factor plane in Fig. 7.25b. Each point represents a single signal condition. On this plane, a distinction is made between those signals differing only in duty factor (df) and fundamental frequency (f_0) (see insert key)[1]. The axes are scaled so that the variances of stimulus values on the two coordinates are equal. Dimension I can be associated with the physical attribute duty factor. Dimension III can be interpreted as fundamental frequency. The signal conditions can be divided according to duty factor and fundamental frequency, as shown by the dashed lines. Considering the direction of subject vectors in Fig. 7.25a, one sees there is a general preference for low duty factor and low fundamental frequency signals.

Multidimensional scaling in its many forms appears particularly promising for quality assessment of speech signals. Synthetic speech is a good case in point. Here the intelligibility may be made high, and the interest is in finding physical factors that relate to (and may be used to improve) naturalness. In other instances, multi-dimensional scaling has been valuable in assessing quality degradations due to non-linear distortions in speech transmission systems.

7.7. Calculating Intelligibility Scores from System Response and Noise Level: The Articulation Index

Articulation tests, properly done to get stable and consistent results, are immensely time consuming. More desirable is the ability to estimate intelligibility from the physical transmission characteristics of the system; for example, from the frequency-amplitude response and the noise level. Under certain restrictive conditions, the well-known articulation index is a technique for making such an estimate (FRENCH and STEINBERG). The concept has been extended and organized into graphical and tabular

[1] Each triangle, for example, represents nine different combinations of modulation rate and harmonic content.

operations for rapid, practical application (BERANEK, 1947, 1954; KRYTER).

The articulation index method is limited to particular distortions in systems using conventional "waveform" transmission. These distortions include relatively smooth variations and limitations in the transmission bandwidth, and the masking of the transmitted signal by ongoing, continuous-spectra noises. Under certain conditions, interference caused by temporally interrupted noise, nonlinear amplitude distortion (peak clipping), and masking by reverberation can be accounted for. In general, however, the technique is not applicable to systems whose transmission bands exhibit many sharp peaks and valleys, to periodic line spectra masking noises, to intermodulation distortions and nonlinearities, and to transmission systems generally of the analysis-synthesis type (that is, where the speech information is coded in terms other than the facsimile waveform).

The technique for calculating the articulation index (*AI*) has been described in detail in many other places. The intent here is simply to recall its principles and, in a brief way, to indicate its applicability and utility. Its calculation is illustrated by the graph in Fig. 7.26 (BERANEK, 1954). This plot shows several spectral densities laid off on a special frequency scale. The frequency scale is similar to the mel (pitch) scale. It is experimentally partitioned into twenty bands that contribute equally to intelligibility. The various spectral densities, or rms sound pressure levels per cycle, show: (a) the threshold of audibility

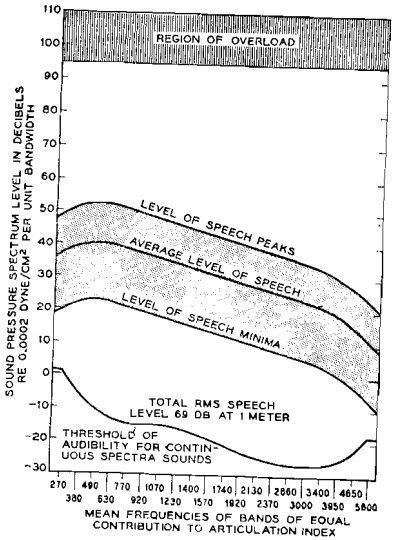


Fig. 7.26. Diagram for calculating the articulation index. (After BERANEK)

for continuous spectra sounds, (b) the peak, average and minimum levels of speech for a man's raised voice at a distance of one meter (see Section 5.17, Chapter V), and (c) an approximate overload spectrum level for the human ear.

In its simplest form, calculation of the articulation index proceeds as follows. The level and shape of the plotted speech spectrum is modified according to the amplification and bandpass characteristics of the transmission system. The spectrum level of any added masking noise is plotted onto the graph. So long as the system response and noise level are such that all of the shaded "speech region" (between minima and maxima) lies above threshold, above the masking noise, and below overload, the intelligibility will be near perfect. In such a case the articulation index is 100%. If any of the speech region is obscured by noise, threshold or overload, the articulation index is diminished by the percentage of the area covered.

Having obtained a number for *AI*, it is necessary to relate it to intelligibility. The relation is an empirical one and is established from articulation tests. As mentioned earlier, articulation tests are subject to considerable variability and their results depend strongly upon testing technique and procedure. Absolute values of scores so derived must be used and interpreted with great discretion. Usually it is more relevant to consider *differences* in intelligibility scores, arrived at by the same technique, than to consider absolute values. Representative empirical relations between intelligibility score and articulation index for a range of test conditions are shown in Fig. 7.27 (KRYTER).

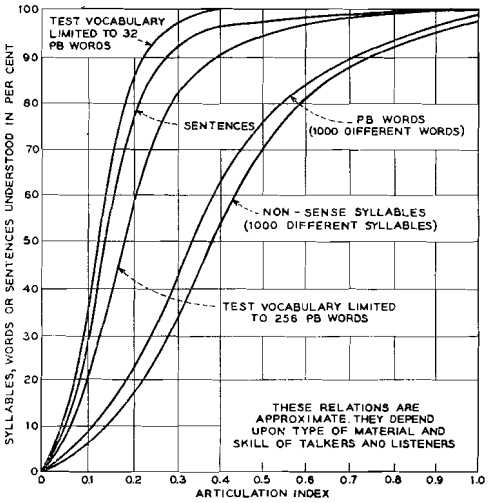


Fig. 7.27. Several experimental relations between articulation index and speech intelligibility. (After KRYTER)

7.8. Supplementary Sensory Channels for Speech Perception

Supplementary methods for speech communication are of great importance to persons either totally deafened or with partial auditory impairment. Not only is it difficult for them to hear the speech of others, but they cannot hear their own speech. It consequently is common that they also experience difficulty in speaking.

At least three avenues have been considered at the research level for providing supplementary perceptual channels and machine aids for speech communication. They include visual, tactile, and auditory approaches. The latter is oriented toward making use of whatever hearing ability may remain. Each approach can be illustrated briefly by a specific example. Other interests and efforts exist in the area.

7.81. Visible Speech Translator

One well-known technique for visually displaying speech information is the "Visible Speech" method (POTTER, KOPP, and GREEN). A real time sound spectrograph, called a Visible Speech Translator, produces a running, continuous spectrographic display on a phosphor screen (RIESZ and SCHOTT; DUDLEY and GRUENZ). The format is similar to the conventional sound spectrogram (shown in Section 5.14, Chapter V) except that the pattern is "painted" continuously, either on a rotating cathode ray tube or on a phosphor belt. As the trace advances with time, a given duration of the past speech is retained and displayed by the persistence of the trace.

Some experiments have been made into the ability of viewers to "read" the direct-translator displays (POTTER, KOPP, and GREEN). The results showed that after relatively lengthy training, trainees were able to converse among themselves by talking clearly and at a fairly slow rate. Within the limits of their vocabulary, they learned to carry on conversations with about the same facility as a similarly advanced class in a foreign language. The learning rates observed in the tests correspond roughly to 350 vocabulary words per one hundred hours of training.

Real-time spectrographic displays appear to have more promise for speech teaching, that is, articulatory training, than for speech reading. Some research has applied spectrographic methods in teaching articulation to deaf children (STARK, CULLEN, and CHASE; RISBERG; PICKETT).

Because of the complex apparatus and important training procedures, visible speech techniques still remain in the realm of research. These and related methods—for example, the display of articulatory data and of formant data—are all valid problems for research and may hold

potential for supplementary communication. Particularly promising are simple devices which signal rudimentary speech features, such as voicing, friction and stop gap (UPTON). At present, however, much remains to be learned about modes of visual presentation of speech information.

7.82. Tactile Vocoder

The sense of touch offers another possibility for real-time communication. A filter bank analyzer, similar to that used in a vocoder, is one means for supplying cutaneous information about the short-time amplitude spectrum of speech (PICKETT, 1969). The technique is shown in Fig. 7.28. Ten contiguous bandpass filters, spanning the frequency range 100 to 8000 cps, receive the speech signal. Their outputs are rectified and smoothed to obtain values of the short-time spectrum at ten frequency positions. The ten time-varying voltages are used to amplitude-modulate individual sinusoidal carriers of 300 cps[1]. The modulated carriers are then applied to fingertip vibrators (actually bone conduction transducers). The analyzing channel of lowest frequency is led to the small finger of the left hand, and the channel of highest frequency connects to the small finger of the right hand.

After practice with the presentation, some subjects are able to make sound discriminations comparable to, and sometimes better than, that

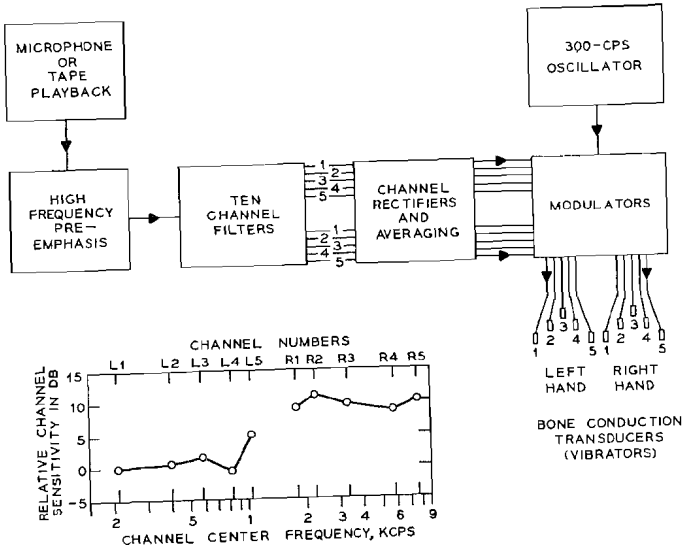


Fig. 7.28. Block diagram of a tactile vocoder. (After PICKETT)

[1] This tactile "carrier" is used because the frequency range of the skin's vibratory sensitivity is limited to about 100 to 800 cps.

achieved in lip reading. When the tactile information is used in combination with lip reading, the ability to identify spoken words is considerably increased. For example, in one measurement of discrimination among 12 words, the lip reading response was about 60% correct. When supplemented by the tactile information, the response increased to 85% (PICKETT).

As in the visible speech method, the vocoder apparatus for tactile display is relatively complex. A much simplified tactile device is shown in Fig. 7.29 (KRINGLEBOTN). This device employs only five vibrators

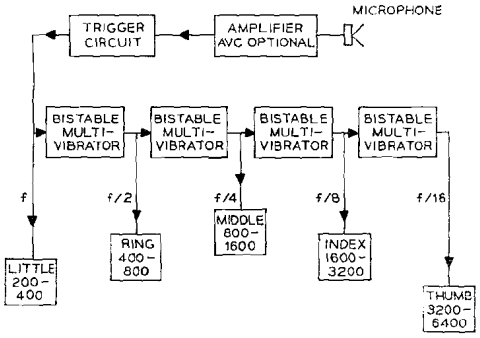


Fig. 7.29. A frequency-dividing tactile vocoder. (After KRINGLEBOTN)

applied to one hand. No filters are used, but stages of frequency division are arranged to divide five frequency ranges of the speech signal so as to vibrate the fingers individually. The vibrations on each finger are felt most strongly in the frequency range 200 to 400 cps. Because of the successive frequency divisions, this sensitivity range corresponds to successively higher frequency ranges in the input signal when distributed over the fingers, going from little finger to thumb. This method probably transmits some frequency information about the speech signal in terms of tactile frequency and other frequency information in terms of tactile location. Training tests with this system have been carried out with deaf children (KRINGLEBOTN).

A number of other efforts in kinesthetic and tactile communication are in progress. Although many of these aim toward machine aids for the blind rather than for the deaf, the presentation of sensory information involves problems common to both areas (BLISS; LINVILL).

7.83. Low Frequency Vocoder

The conventional electronic hearing aid is an amplifying and frequency shaping device. It facilitates the use of whatever residual

hearing a deafened person may have. In severe cases, however, the residual hearing is often confined to a very small bandwidth, usually at the low-frequency end of the audible spectrum. For example, a typical audiogram might show 60 to 80 db loss from 30 to 400 cps and 110 db above 500 cps.

One proposal is to make maximal use of such residual hearing. Slowly varying signals that describe the short-time speech spectrum (such as vocoder channel signals) are modulated either onto sinusoidal carriers of very low frequency, or onto distinctive complex signals of relatively small bandwith (PIMONOW). In one implementation, seven spectrum channels extending to 7000 cps are used. The rectified, smoothed outputs amplitude modulate the same number of low-frequency, sinusoidal carriers. The carriers are spaced from 30 to 300 cps. The modulated carriers are summed and presented as an auditory signal. In an alternative arrangement, the modulated signals are non-sinusoidal and include a low-frequency noise band, a periodic pulse train, and a band of actual speech. In one series of experiments, deafened subjects who could not use ordinary hearing aids apparently learned to discriminate well among a limited ensemble of words (PIMONOW).

Various devices for spectrum shifting, transposing or dividing have also been considered (JOHANSSON; GUTTMAN, and NELSON; LEVITT and NELSON). These devices generally aim to recode high-frequency information into a lower-frequency range where residual hearing exists. Like visible speech displays, their value appears to lie more in articulatory training than in speech reception. Like the other sensory aids discussed in this section, frequency scaling devices are still in the research stage. Extended experimentation and technical development will determine their potential as practicable aids to hearing.

VIII. Systems for Analysis-Synthesis Telephony

The discussions in Chapters III and IV considered the basic physics of the mechanisms for speech production and hearing. The topics of Chapters V, VI and VII set forward certain principles relating to the analysis, artificial generation, and perception of speech. The present and final chapter proposes to indicate how the foregoing results, in combination, may be applied to the efficient transmission of speech.

Efficient communication suggests transmission of the minimum information necessary to specify a speech event and to evoke a desired response. Implicit is the notion that the message ensemble contains only the sounds of human speech. No other signals are relevant. The basic problem is to design a system so that it transmits with maximum efficiency only the perceptually significant information of speech.